



```
'SEARCH.BAS 11/21/94

'Finds pixels with intensities greater than fmin
'Calculates the match score of bright pixels

'The input file eightmer.dat is based on fs8mer.exe
'The output file score.dat lists the bright pixels
' and gives the number of 1 and 2 mismatch-related
' pixels

DIM a$(1000), f%(1000), m1%(1000), m2%(1000)

tstart = TIMER
inf$ = "eightmer.dat"
outf$ = "score.dat"
OPEN inf$ FOR INPUT AS #1
OPEN outf$ FOR OUTPUT AS #2
CLS

'Read the input file and store the bright pixels
fmin = 90 'threshold value for inclusion
n = 0: fmax = 0
WHILE NOT EOF(1)
LINE INPUT #1, g$
seq$ = MID$(g$, 1, 8)
intens = VAL(MID$(g$, 9, 6))
IF intens >= 90 THEN
n = n + 1
a$(n) = seq$
f%(n) = intens
IF intens > fmax THEN fmax = intens
PRINT n;
END IF
WEND
PRINT
PRINT USING "#### intensity values above ####"; n; fmin
PRINT USING "Highest intensity is ####"; fmax

'Calculate m1 and m2 for each bright pixel
' m1 and m2 are the number of other pixels that are related
' by 1 and 2 mismatches, respectively

m1max = 0 'Keep track of highest m1 score
FOR j = 1 TO n
PRINT j;
FOR i = 1 TO n
m = 0
FOR k = 1 TO 8
IF MID$(a$(j), k, 1) <> MID$(a$(i), k, 1) THEN m = m + 1
NEXT k
IF m = 1 THEN m1%(j) = m1%(j) + 1
IF m = 2 THEN m2%(j) = m2%(j) + 1
NEXT i
IF m1%(j) > m1max THEN m1max = m1%(j)
NEXT j

PRINT #2, USING "SEARCH.BAS & "; DATE$; TIME$
PRINT #2, USING "Input file: & Output file: &"; inf$; outf$
PRINT #2, USING "#### intensity values above ####"; n; fmin
PRINT #2, USING "Highest intensity is ####"; fmax
```

```

PRINT #2, USING "Greatest number of 1-mismatch relations is ##"; mlmax
PRINT #2,
PRINT #2, "List of probes with highest intensity and best matching"
PRINT #2, " f      m1  m2      sequence"
  \ k = 1 TO n
IF f%(k) > .4 * fmax AND m1%(k) > .4 * mlmax THEN
  PRINT #2, USING "#### ## ## &"; f%(k); m1%(k); m2%(k); a$(k)
END IF
NEXT k
PRINT #2, CHR$(12)

'Sort according to f
s% = n \ 2
DO WHILE s% > 0
  FOR i% = s% TO n - 1
    j% = i% - s% + 1
    FOR j% = (i% - s% + 1) TO 1 STEP -s%
      IF f%(j%) >= f%(j% + s%) THEN EXIT FOR
      SWAP f%(j%), f%(j% + s%)
      SWAP m1%(j%), m1%(j% + s%)
      SWAP m2%(j%), m2%(j% + s%)
      SWAP a$(j%), a$(j% + s%)
    NEXT j%
  NEXT i%
  s% = s% \ 2
LOOP
PRINT #2,
PRINT #2, " f      m1  m2      sequence"
FOR k = 1 TO n
  PRINT #2, USING "#### ## ## &"; f%(k); m1%(k); m2%(k); a$(k)
NEXT k

PRINT CHR$(12)
'Sort according to m1
s% = n \ 2
DO WHILE s% > 0
  FOR i% = s% TO n - 1
    j% = i% - s% + 1
    FOR j% = (i% - s% + 1) TO 1 STEP -s%
      IF m1%(j%) >= m1%(j% + s%) THEN EXIT FOR
      SWAP f%(j%), f%(j% + s%)
      SWAP m1%(j%), m1%(j% + s%)
      SWAP m2%(j%), m2%(j% + s%)
      SWAP a$(j%), a$(j% + s%)
    NEXT j%
  NEXT i%
  s% = s% \ 2
LOOP
PRINT #2,
PRINT #2, " f      m1  m2      sequence"
FOR k = 1 TO n
  PRINT #2, USING "#### ## ## &"; f%(k); m1%(k); m2%(k); a$(k)
NEXT k

PRINT USING "Time= ####.## seconds"; TIMER - tstart

```

```

'CONSENS.BAS 1/8/95
'Derive a consensus sequence from the highest scoring probes

DIM a$(70), m%(1, 70, 70), f(70), s(-20 TO 20, 4)
CLS
INPUT "Input file: ", inf$
INPUT "Output file: ", outf$
OPEN inf$ FOR INPUT AS #1
OPEN outf$ FOR OUTPUT AS #2

LINE INPUT #1, descr$      'File description
INPUT #1, pl 'Probe length
INPUT #1, n  'Number of sequences
FOR j = 1 TO n
LINE INPUT #1, a$(j)
NEXT j
CLOSE #1

'Initialize the mismatch matrix
FOR z = 0 TO 1: FOR i = 1 TO n: FOR j = 1 TO n
m%(z, i, j) = 100
NEXT j: NEXT i: NEXT z

PRINT #2,
PRINT #2, "CONSENS.BAS "; DATE$; " "; TIME$
PRINT #2, : PRINT #2,
PRINT #2, "Input file: "; inf$; " Output file: "; outf$
PRINT #2, descr$
PRINT #2, USING "The ## ##-mer sequences with the highest scores are:";
n; pl
PRINT #2,
FOR j = 1 TO n
PRINT #2, USING "## &"; j; a$(j)
NEXT j
PRINT #2, : PRINT #2,

z = 0
PRINT #2, USING "z=##"; z
PRINT #2, " ";
FOR k = 1 TO n: PRINT #2, USING "##"; k; : NEXT k
FOR i = 1 TO n
PRINT #2,
PRINT #2, USING "## "; i;
FOR j = 1 TO n
m = 0
FOR k = 1 TO pl
IF MID$(a$(j), k, 1) <> MID$(a$(i), k, 1) THEN m = m + 1
NEXT k
m%(0, i, j) = m
IF m <= 2 THEN PRINT #2, USING " ##"; m; ELSE PRINT #2, " .";
NEXT j
NEXT i
PRINT #2, : PRINT #2,

z = 1
PRINT #2, USING "z=##";
PRINT #2, " ";

```

```

FOR k = 1 TO n: PRINT #2, USING "##"; k; : NEXT k
FOR i = 1 TO n
PRINT #2,
PRINT #2, USING "## "; i;
FOR j = 1 TO n
    m = 0
    FOR k = 1 TO p1 - 1
        IF MID$(a$(j), k, 1) <> MID$(a$(i), k + 1, 1) THEN m = m + 1
    NEXT k
    m%(1, i, j) = m
    IF m <= 2 THEN PRINT #2, USING " #"; m; ELSE PRINT #2, " .";
NEXT j
NEXT i

PRINT #2, : PRINT #2,

'Mark all sequences with a 100 tag
FOR i = 1 TO n: f(i) = 100: NEXT i
'Designate the first sequence as the origin
f(1) = 0

'Find the frames of sequences that can be aligned
FOR i = 1 TO n
FOR j = 1 TO n
IF m%(1, i, j) <= 2 AND f(i) <> 100 THEN
    f(j) = f(i) + 1
END IF
NEXT j
NEXT i

FOR i = 1 TO n
FOR j = 1 TO n
IF m%(1, j, i) <= 2 AND f(i) <> 100 THEN
    f(j) = f(i) - 1
END IF
NEXT j
NEXT i

FOR i = 1 TO n
FOR j = i + 1 TO n
IF m%(0, i, j) <= 1 AND f(i) <> 100 THEN
    f(j) = f(i)
END IF
NEXT j
NEXT i

PRINT #2, : PRINT #2,
PRINT #2, "Alignment criteria: <=1 mismatch allowed for z=0"
PRINT #2, "    <=2 mismatches for z=1"

PRINT #2,
PRINT #2, "The aligned sequences are:"
'Print the aligned sequences
FOR i = 1 TO n
IF f(i) <> 100 THEN
    PRINT #2, SPACES$(15 + f(i)); a$(i)
END IF

```

```

NEXT i
PRINT #2, : PRINT #2,

'Accumulate the sequence scores
offset = 0
FOR i = 1 TO n
IF f(i) <> 100 THEN
    FOR k = 1 TO pl
        g = INSTR("ACGT", MID$(a$(i), k, 1))
        s(offset + k + f(i), g) = s(offset + k + f(i), g) + 1
        'PRINT offset + k + f(i); g; "        ";
    NEXT k
END IF
NEXT i

PRINT #2, CHR$(12)
PRINT #2, "CONSENS.BAS "; DATE$; " "; TIME$
PRINT #2, USING "Input file: &      Output file: &"; inf$; outf$
PRINT #2, USING "### ##mer sequences"; n; pl
PRINT #2, descr$
PRINT #2,

PRINT #2, "The frequencies of bases in the aligned sequences are:"
PRINT #2,
'Print the scores
FOR g = 1 TO 4
FOR j = -10 TO 18
PRINT #2, USING "## "; s(j, g);
'PRINT USING "## "; s(j, g);
NEXT j
PRINT #2,
NEXT g

'Find and print the consensus
c$(0) = "-": c$(1) = "A": c$(2) = "C": c$(3) = "G": c$(4) = "T"
FOR j = -10 TO 18
most = 0: mg = 0: sum = 0: b$ = "-"
FOR g = 1 TO 4
IF s(j, g) > most THEN most = s(j, g): mg = g
sum = sum + s(j, g)
NEXT g

'A base is defined if present in at least 2 sequences
' and 55% of those aligned at that position

IF most >= 3 THEN
    IF most / sum > .5 THEN b$ = c$(mg)
END IF
PRINT #2, USING " & "; b$;
cons$ = cons$ + b$
NEXT j
PRINT #2, : PRINT #2, : PRINT #2, "The consensus sequence is: "; cons$:
PRINT #2,
PRINT cons$

PRINT #2, : PRINT #2,
PRINT #2, "The correct sequence is TCAACATCACCTACCA"

```

```
PRINT #2,  
PRINT #2, "The stray sequences are:"  
FOR i = 1 TO n  
IF f(i) = 100 THEN PRINT #2, SPACE$(5); a$(i)  
NEXT i
```